Department for Transport

TAG UNIT M1.2

Data Sources and Surveys

May 2024

Department for Transport

Transport Analysis Guidance (TAG)

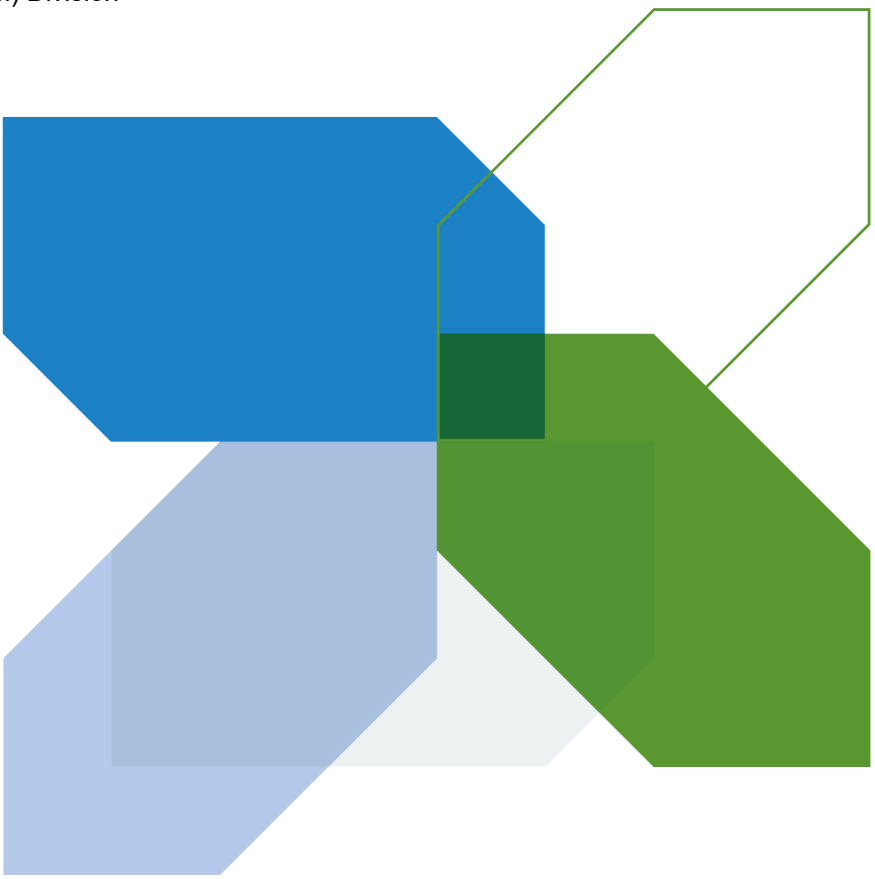
https://www.gov.uk/transport-analysis-guidance-tag

This TAG Unit is guidance for the **Modelling Practitioner**

This TAG Unit is part of the family **M1 - Modelling Principles**

Technical queries and comments on this TAG Unit should be referred to:

Transport Appraisal and Strategic Modelling (TASM) Division



Department for Transport
Zone 2/25 Great Minster House
33 Horseferry Road
London
SW1P 4DR
tasm@dft.gov.uk

Contents

1. Introduction

1.1 Introduction to Data Sources

1.1.1 The aim of this unit is to identify what sources of transport data are available to practitioners developing transport models. The unit also addresses the methods used for gathering data including survey methodology.

1.1.2 Data collection is necessary to build a transport model and to inform the numerical parameters used in the model, many of which cannot be observed directly, and therefore need to be estimated using a sample of data (**model calibration**). Models should be compared against observed data for the current or a recent year, which has been collected independently from the data that is used to calibrate the model parameters, to demonstrate that the model is of sufficiently good quality and robust (**model validation**) to form the basis from which to calculate future forecasts. For further details on model calibration and validation see TAG units M3.1 and M3.2 in the Guidance for the Modelling Practitioner.

1.1.3 One of the main constraints on data collection is the cost of conducting transport surveys. This cost can be reduced by minimising the need to collect new data. In order to reduce the costs of building a transport model, practitioners should invest time in the investigation of all relevant and available transport data sets and their suitability for the model being developed.

1.1.4 This unit provides comprehensive guidance on existing data sources, which are in general the major available data sets and describes the data which are contained within them. It addresses transport data which are relevant to the development of demand models as well as supply or assignment models, but the list is not exhaustive nor covers additional data which may be available locally. Practitioners should consider contacting sub-national transport bodies or local authorities to understand availability and access to local data sources.

1.1.5 This unit also contains advice on conducting study specific surveys. It addresses the requirements of each type of survey and discusses the standards to be expected with each survey type including survey accuracy.

1.1.6 The methods and data described in this unit are non-exhaustive, and new data sources and uses will no doubt continue to emerge. The principles set out in this unit for understanding the merits and addressing the limitations of data sources should be followed for these as well.

1.2 Relationship of this Unit to Other Advice

1.2.1 This unit is a companion to the following TAG units:

- TAG Unit M1.1 Principles of Modelling and Forecasting
- TAG Unit M2.1 Variable Demand Modelling
- TAG Unit M3.1 Highway Assignment Modelling
- TAG Unit M3.2 Public Transport Assignment Modelling

1.2.2 This unit excludes advice on processing traffic data for the development of demand matrices for transport models, which is set out in TAG Unit M2.2 Base Year Demand Matrix Development. Advice contained therein should be referred to before finalisation of the data assembly planning (see section 2.2).

1.2.3 The use of a land-use / transport interaction (LUTI) model, which provides 'planning data' inputs that interact with a transport model, requires a specialised set of data requirements that are not considered here. For advice on LUTI models, see Supplementary Guidance – Land Use/Transport Interaction Models.

1.3 Structure of this Unit

1.3.1 After this introductory section, this unit has been divided into four main sections.

1.3.2 Section 2 discusses considerations when collating and collecting data to develop transport models, which contains guidance on:

- planning for data assembly
- accuracy and suitability of data, and principles for treatment of limitations
- general considerations for collecting data

1.3.3 Section 3 contains guidance on available data sets that could be used for transport modelling and includes:

- data from existing models
- land use and planning data
- geographical data
- network data
- tracking data
- ticket data
- non-intrusive surveys
- interview surveys

1.3.4 Section 4 contains guidance on conducting surveys for specific studies and includes:

- data specification and requirements for demand models
- conducting highway traffic surveys and the standards and tolerances required for the data
- public transport surveys and the methods to be used for gathering these data

1.3.5 Section 5 contains a summary of guidance on modelling purposes that different data sources may be used for.

2. Data Considerations

2.1 Introduction

2.1.1 Data assembly is often one of the most resource-intensive aspects of transport modelling. It is therefore essential to carefully plan this step to minimise project risks and use resources efficiently. Development of a proportionate data assembly plan requires a thorough understanding of existing data sources, their suitability for planned use(s), and gaps in the data that should be filled through data collection.

2.1.2 This stage also includes documentation of the data sources, geographical and temporal coverage, possible biases and measures proposed to correct them. This will form part of the Model Specification Report (see TAG Unit M1.1 Principles of Modelling and Forecasting). A separate data and surveys report needs to be prepared as part of model development (see the Guidance for the Technical Project Manager).

2.1.3 This section provides advice on planning for the data assembly process, assessment of accuracy and suitability of data sources, and key considerations in developing the data collection specification.

2.2 Planning for Data Assembly

2.2.1 Planning for data assembly should form part of an overall planning stage for any major modelling task that also considers rationale for modelling, specific objectives, and specification of a suitably proportionate approach. Practitioners should refer to TAG Unit M1.1 Principles of Modelling and Forecasting for general guidance on the rationale for modelling and the considerations for defining the overall structure of the model, linked to specific model objectives, and hence the data requirements.

2.2.2 The extent of the data required will vary depending on the study. For example, a full model development will require many of the data sources outlined in TAG Unit M1.1 Principles of Modelling and Forecasting, whereas a recalibration task may just require new traffic counts. Consideration should also be given to the potential longer-term use of the proposed model and whether there may be requirements to extend the coverage or detail of the model in the near future, which in turn may require additional survey data.

2.2.3 Following the identification of the data required, existing data sources (see Section 3) should be identified and reviewed to establish their credibility and accuracy, and their suitability for the intended modelling task. This should be based on an investigation of possible biases and inconsistencies in the data, and whether these can be remedied (see Section 2.3).

2.2.4 At this stage, understanding availability and taking advantage of existing transport data can be of vital importance in the specification and implementation stages of model development, as referred to in the Guidance for the Technical Project Manager, and will reduce costs.

2.2.5 When existing data are expected to be used, sufficient time must be allowed for them to be identified, obtained from their current custodian, reprocessed as necessary, and checked for consistency and validity. Further delays may be incurred if these checks reveal that the data cannot be used as intended.

2.2.6 Finally, after existing data sources and their suitability are reviewed, gaps in the existing data should be identified and requirements for any additional data collection should be established. Practicality of the collection of this bespoke data, the costs and the time required to collect should be addressed. This bespoke data collection requires specification and should be outlined as part of the intended model development approach in Model Specification Reports (see TAG Unit M1.1 Principles of Modelling and Forecasting).

2.3 Accuracy and Suitability of Data

2.3.1 There are several factors that determine quality of data sources and whether they are suitable for their planned use, these include:

- the provenance of the data source
- availability and transparency of reports describing the nature of the data, and the collection and processing steps
- definitional issues that might impact on compatibility with other sources
- suspected or known biases, their size, and how these have been treated
- spatial coverage and date of data collection and/or release
- any restrictions on access or use of data

- age of data, particularly in locations that have seen substantial changes since their collection

2.3.2 In general, errors in the data are influenced by data collection, sampling approaches, and data processing methods. These include measurement errors caused by malfunctions or calibration of data collection equipment, human errors made during data processing, and statistical errors in the processed data.

2.3.3 When an estimate of a quantity such as traffic flow has been made from a sample of data, it is desirable to know not only the estimated value but also how reliable this estimate is. These estimates are subject to statistical errors, reflecting the difference between the estimated and the true value of the quantity. A convenient way of expressing the statistical error is to state limits which, with a given probability (usually 95%) include the true value. It is then possible to state, for example, that the true value is unlikely to exceed some upper limit, or to be less than a lower limit, or to lie outside a pair of limits. This information may be more important than the estimate itself. These limits are known as "confidence limits", i.e., they are limits within which it can be stated, with a given degree of confidence, that the true value lies (TAG Unit M2.2 Base Year Matrix Development demonstrates the calculation of these statistical errors for certain data sources).

2.3.4 The adoption of a particular confidence interval implies a decision concerning the accuracy required of the information presented. For example, a 95% confidence interval accepts the chance that the true value will lie outside the given limits only 5 times in 100 occurrences.

2.3.5 Commercial data sources, such as Mobile Network Data (MND), Global Positioning System (GPS) tracking data, and ticketing data, are generally not collected specifically to obtain input data for transport models. They are usually processed using methods that are subject to commercial confidentiality; therefore, details of the initial stages of data processing may not be known to practitioners. Hence, the statistical errors in the data are difficult to quantify.

2.3.6 Nevertheless, it is important to ask data providers to provide as much information as possible about processing assumptions, sample size, definitional inconsistencies, and gaps in the data so that strengths and limitations of the data are understood. This is important to ensure that the data are used appropriately. Further advice on the use of MND and other demand data is provided in TAG Unit M2.2 Base Year Demand Matrix Development. MND may be suitable for purposes in addition to demand matrix development described in this unit; these should be established by practitioners through dialogue with MND data providers.

2.3.7 In addition to sources of error in the data discussed above, many data sources may include biases. Biases refer to systematic differences between the estimates and true values, which are linked with some characteristics of the data and tend to move the estimates towards a specific direction with a certain magnitude. This contrasts with statistical errors, that tend to be randomly distributed and are expected to have no overall net effect on mean of all observations in the data.

2.3.8 During the investigation stage of the data assembly process, known and suspected biases in the existing data sources should be identified, their size and direction should be established, and either the data should be discarded or an appropriate approach to treat the biases should be specified. When collecting new data, the data collection methodology should carefully consider methods to minimise biases in the data. Any residual biases should then be addressed during the data processing steps.

2.3.9 Validity of data is a key consideration in assessing whether existing data sources are suitable for intended use(s). One factor affecting data validity is age. A single threshold should not be used to determine data suitability. Instead, age should be considered together with the extent to which population, land use, transport network, and travel demand patterns have changed, as well as the quality and consistency of documentation that describes data collection and processing steps.

2.3.10 For example, significant changes post data collection, such as the opening of a new bypass, a major new development or changes in public transport services in the vicinity of model internal area, would make the use of data collected prior to these data inappropriate. In other cases, older data may be acceptable where there has been little change to land use and transport network, however, suitable adjustments may be necessary to account for such changes. Therefore, practitioners should establish evidence on scale of changes to land use and demographic characteristics, transport networks, and travel patterns, and use this evidence to assess the validity of 'old' data sources for the intended use(s). GIS analysis of network mapping data, local authority planning data, and long-term traffic counts are examples of data sources that could be analysed to provide such evidence.

2.3.11 Based on the accuracy of the data sources, the extent of identified errors and biases, and the approach to treating them, the suitability of data sources for their planned use(s) should be reviewed. Practitioners should critically consider the data accuracy and bias to first judge whether the data are of sufficient quality to use. The variance of the data errors should be compared with the validation acceptability criteria set out in TAG units M3.1 and M3.2 in the Guidance for the Modelling Practitioner (which predominantly reflect model errors). If the data sampling errors are larger than these criteria, then the data cannot be used to test model performance against these standards. Methods to correct for identified biases will need to be planned and reported.

2.4 Data Collection Considerations

2.4.1 When existing data are deemed to be insufficient or inappropriate, new data may need to be collected.

2.4.2 The key considerations that should be given when new data collection is specified include:

- spatial coverage

- data collection period
- sample size and rate
- the need to represent characteristics of the larger population
- consistency of definitions with existing data sources
- method of data collection
- data privacy

2.4.3 Data collection is often the largest cost of building a transport model. The most cost effective way to manage this is by minimising the quantity of new surveys. The best way to minimise data collection requirements is by taking the data needs into account in the design of the model. Most models split the spatial area to be studied into sectors and are further sub-divided into zones. Data collection can be reduced by careful design of the sectors and zones (for example, survey locations might follow zone boundaries that are designed with a minimum number of crossing points, such as following rivers, railways, or major roads). For guidance on zonal design, see TAG units M3.1 and M3.2 in the Guidance for the Modelling Practitioner.

2.4.4 Depending on the purpose of the model being developed, surveys should typically be carried out during a 'neutral', or representative, period. The period(s) selected will depend on the model purpose and may for example, represent average working weekdays, or a Saturday. Normally the period should not be affected by conditions or irregular events that may abnormally affect travel demand patterns. However, there can be instances where the purpose of the model necessitates demand to be representative of a particular period that is different from neutral period defined above, for example, school holidays, or holiday periods in regions with relatively high levels of seasonal tourism, or in planning for specific events. The period for the surveys should be selected with careful consideration of the purpose of the transport model, ensuring that the conditions being surveyed (e.g. traffic flow) are representative of the transport condition being analysed/modelled.

2.4.5 When the purpose of the model is to represent neutral conditions, practitioners are advised to gather initial evidence (for example, long-term count data) to understand current travel patterns and help determine the most appropriate neutral periods for their surveys. In many cases, surveys can be carried out any time of the year on any given day of the week that travel patterns are considered to be representative of the neutral condition being modelled, provided that:

- weather conditions are unlikely to influence travel demand being surveyed
- there are no special events that are likely to influence travel demand being surveyed

- adequate lighting is available, where the survey method may have safety considerations, or visibility affects data quality

- the weeks before/after Easter and Christmas holidays, the Thursday and/or Friday before and all of the week of a bank holiday, and the school holidays are avoided as demand is likely to be different than neutral conditions

2.4.6 Transport models typically model time periods of the day that represent neutral conditions. Practitioners should refer to TAG unit M2.1 Variable Demand Modelling for guidance on designing the modelled time periods to represent the variation in travel conditions across different hours within the modelled day. For appraisal purposes, annualisation factors are then used to expand the benefits estimated from the modelled hours to cover the whole day and to represent all impacts over a full year. TAG unit A1.3 User and Provider Impacts describes the need to use annualisation factors and summarises the key principles that should be considered when they are calculated. Methods to calculate annualisation factors are well established and examples of good practice are documented and available that practitioners can refer to[1].

2.4.7 There are specific considerations that practitioners should apply when planning for data collection and assembly for use to calculate annualisation factors. These include:

- adequate local data collected from traffic or travel surveys to allow for annualisation factors to be derived for appraisal purposes.

- annual data sources to include seasonal variations that provide the basis to directly expand model outputs to annual estimates. Examples include long-term traffic counts, public transport ticketing data, or mobile network data. An acceptable exception is when there is confidence that the period over which data are collected (e.g. historic neutral months) is a reasonable representation of annual average condition.

- the number and location of long-term traffic count sites may depend on the purpose for which the factors are derived. For example, for a scheme that is about delivering improvements in journey time on the inter-urban network, and other impacts are relatively minor, the profile of flow along the main corridor will be critical, and may be the only consideration, in which case a single long-term count location could be adequate. However, for a scheme impacting air quality and congestion in an urban area, there will be quite distinct patterns of changes and profiles of demand between road types and locations, which the model needs to capture, in which case it may be necessary to obtain data from a range of count sites to analyse variations in the patterns of demand.

- it is important that the accuracy of local data used to derive annualisation factors is understood, and the data are judged to have sufficient quality (see Section 2.3 for advice on accuracy and suitability of data). This can be

[1] For example, see A303 Project Team and National Highways (2018), AECOM and PBA (2020), Atkins (2018), WSP (2020)

particularly important for appraisal results. The judgment should be applied by practitioners based on factors such as the intended use of the model, nature and materiality of the impacts, and an understanding of temporal variations in network performance.

- it is likely that demand and supply in non-modelled periods differ by mode, for example, there is usually limited public transport provision overnight, and hence separate consideration should be given to the relevant travel modes.

2.4.8 Methods for data collection can generally be divided into four categories, listed in ascending order of likely cost:

- non-intrusive automated methods (for example automated traffic counts; emergent data sources including data from electronic devices may fall into this category)
- non-intrusive manual methods (for example, traffic counts or manual turning count surveys collected by an observer, or observed journey time surveys)
- intrusive automated methods (collecting data from installed mobile applications or GPS sensors installed in a sample of vehicles)
- intrusive manual methods (interviews with potential customers, which include roadside interviews and travel diary surveys)

2.4.9 Where there is an option, the top priority is to minimise the reliance on intrusive manual data collection methods. This is likely to give the greatest cost saving, although some travel information (e.g. trip purpose or person characteristics) cannot normally be gathered without some form of intrusive manual surveys.

2.4.10 Section 4 contains guidance on conducting bespoke surveys for specific studies. TAG Unit M2.2 Base Year Demand Matrix Development provides further guidance on specific considerations that should be given to collecting demand matrix data.

3. Existing Data Sources

3.1 Introduction

3.1.1 This section identifies existing sources of data that may be useful in transport modelling. It describes each data source together with its strengths and limitations, and where to find the data.

3.1.2 This section does not provide an exhaustive list of data sources that can be used. When considering these or other data sources, the practitioners should follow the general advice in Section 2 to ensure the data source is suited to the task, with sufficient accuracy.

3.1.3 This section covers existing data sources in the following categories:

- **Data from existing models** that include the National Trip End Model (NTEM), National Road Traffic Projections (NRTP), MOIRA (models to forecast rail demand and revenue), and the National Atmospheric Emissions Inventory (NAEI). Base year demand matrices from existing transport models can also be used.

- **Land use and planning data**, which include population, employment and labour market data from the Office for National Statistics (ONS), Valuation Office Agency (VOA) floorspace, Department for Education school census, Driver and Vehicle Licensing Agency (DVLA) vehicle licensing statistics, Mosaic geodemographic and Acorn geodemographic data, the Index of Multiple Deprivation, and AddressBase.

- **Geographical and network data**, including data from Ordnance Survey (OS), such as MasterMap's topography layer and terrain mapping, MasterMap highways network, and open roads. It also lists ONS boundary data, OpenStreetMap, the Major Roads Database, Openrouteservice, the National Cycle Network and National Rail Timetable, CIF railway timetable, Traveline National Dataset, Realtime trains, Department for Transport's Bus Open Data Service (BODS) and Maritime Statistics, Office of Rail and Road (ORR) station usage data, the National Rail Enquiries open data feed, National Public Transport Access Nodes (NaPTAN).

- **Tracking data**, including mobile network data and GPS tracking data.

- **Ticket data** from Electronic Ticket Machine (ETM), Latest Earnings Networked Nationally Overnight (LENNON), and the RUDD Rail Usage and Drivers Dataset

- **Data from non-intrusive surveys**, such as Department for Transport's Rail Statistics, Road Traffic Statistics, Congestion Statistics, as well as National Highway's Traffic Information System (WebTRIS). Other sources of data are the Trip Rate Information Computer System (TRICS) and data from local authorities. Video analysis may also be used to collect traffic data.

- **Data from interview surveys**, including the National Travel Survey (NTS), Census, the Civil Aviation Authority airport passengers survey, the International Passenger Survey, the Continuing Survey of Road Goods Transport, the National Travel Attitudes Survey, the National Highways and Transport Network (NHT) Public Satisfaction Survey, the National Rail Passenger Survey, the Department for Transport's Bus Statistics, the Bus Passenger Survey, the Strategic Roads User Survey, the ONS Living Costs and Food Survey, the Time Use Survey, the Annual Survey of Hours and Earnings, and the Business Register and Employment Survey (BRES). Other

sources include local household interview surveys, and roadside interview surveys.

3.2 Data from Existing Models

National Trip End Model

3.2.1 The National Trip End Model (NTEM) includes forecasts of population, households, workforce, and jobs for up to 2061, based on observed values from 2011 Census. These are used in a series of models to predict employment, car ownership, trip ends, and overall traffic growth. NTEM is owned and updated by the Department for Transport. The outputs from NTEM can be viewed in the Trip End Model Presentation Program (TEMPro), which is free to download from the Department's website. This includes the TEMPro data report and various other documentations which should be read prior to using TEMPro.

3.2.2 TEMPro is designed to allow detailed analysis of pre-processed trip-end, car ownership, and population/workforce planning data from the NTEM. The pre-processed data are themselves the output from a series of models developed and run by the Department, which are included in the NTEM database.

3.2.3 The current version is multi-modal, providing data on trips on foot, by bicycle, motor vehicle (both as a driver and passenger), by rail, and by bus. Users should note, however, that TEMPro trip ends by mode are based on average rates over a wide area and do not necessarily take into account the accessibility of each zone by each transport mode. Any such multi-modal segmentation at fine spatial level of detail should be accompanied by a narrative on why this is acceptable for a specific application. The trip ends cover personal travel in Great Britain. International and freight trips are excluded from the data.

3.2.4 The NTEM data are available by Middle Layer Super Output Area (MSOA). However, the person trip rates used to forecast changes in trips are based on averages over a wide area and do not necessarily reflect local differences in trip generation patterns. TEMPro estimates of trip ends at any level below aggregate regions (e.g. MSOA, district, or county level) are subject to uncertainty and should not be used as constraints without local verification and possible adjustments.

3.2.5 Where there is particular value in applying trip end functions at a zonal level, such as in land use transport interaction models, CTripEnd (the trip end model program within NTEM) allows practitioners to prepare trip end forecasts for a set of user-defined zones. It may be possible to update zonal population and land use data within NTEM, if more reliable local information is available. This will ensure that the trip ends generated by the software reflect the underlying land use of each zone in detail and that outputs remain consistent with NTEM assumptions at the more aggregate spatial level.

National Road Traffic Projections

3.2.6 The National Road Traffic Projections (NRTP), previously named the National Road Traffic Forecasts (NRTF), are reports presenting forecasts of road traffic, congestion, and emissions for England and Wales, including goods vehicle forecast growth and speed changes across the highway network. The data (trips and miles travelled) cover the Department's Common Analytical Scenarios which are produced using the National Transport Model (NTMv2r). The forecasts (available on NRTP website) are provided by vehicle type, road type, and 'regions' across England.

MOIRA

3.2.7 The MOIRA models (MOIRA and MOIRA2) are single mode models designed to forecast changes in rail passenger loadings, demand and revenue arising from changes to rail timetables. MOIRA captures rail demand and revenue changes caused by changes in three endogenous variables: journey time, frequency and provision of through services. MOIRA2 can also (optionally) capture changes caused by crowding. Results from the MOIRA models include train loadings, origin-destination demands, revenue, journey times, and average fares. The results can also be grouped to Train Operating Companies (TOCs) and regions across England (former Government Office Regions). The MOIRA models are available from the Rail Delivery Group (RDG) and the Department for Transport, with a licence.

3.2.8 The base demand and revenue data in MOIRA are sourced from Latest Earnings Networked Nationally Overnight (LENNON), with adjustments to allocate zonal and rover tickets to specific station pairs. LENNON is based on ticket sales data and hence does not capture end-to-end journeys, only station-to-station journeys. Trip data from LENNON is stored in MOIRA in production/attraction format, under the assumption that the production end (the starting point for a single trip or outward leg of a return trip) is the point of sale of the ticket and hence there are likely to be definitional inconsistencies relating to journey directionality (see paragraph 3.6.9 for more details).

3.2.9 MOIRA may be used to sense check outputs from other transport models to see if they yield a similar interpretation to MOIRA outputs. However, MOIRA should not be used to validate other models.

Origin and Destination Matrix from Rail Data Marketplace

3.2.10 Origin and destination matrix (ODM) contains journeys between all pairs of mainline stations in Great Britain, which can be accessed through the Rail Data Marketplace. The estimated numbers of journeys are produced annually using the MOIRA2.2 base revenue and journeys matrix, which is derived from local ticketing data and Latest Earnings Networked Nationally Overnight (LENNON), which is discussed further in section 3.6.

National Atmospheric Emissions Inventory

3.2.11 The National Atmospheric Emissions Inventory (NAEI) provides annual emissions data by vehicle type for the UK and forecasts vehicle fleet composition until 2035. The fleet composition data are provided by road type for each country in the UK. The projected fleet compositions are consistent with the Department for Transport's National Road Traffic Projections (NRTP).

Existing Base Year Demand Matrices

3.2.12 Transport models developed based on the advice from TAG may already exist and can be used as a data source for other modelling purposes when their geographical coverage, dimensions, strengths and weaknesses are fully understood. For example, National Highways' Regional Traffic Models (RTMs) have wide geographic coverage and therefore demand matrices from them may be suitable for use in other models for more local developments.

3.2.13 Base year demand matrices sourced from existing models usually require specific targeted refinements (e.g. using matrix estimation) by practitioners to meet their own modelling requirements. It is therefore important that the processing steps and assumptions used to develop them are fully understood. Practitioners should obtain the base year 'prior matrices', not the matrices that have gone through matrix estimation.

3.2.14 The factors to consider before using existing demand matrices are outlined in TAG Unit M2.2 Base Year Demand Matrix Development.

3.3 Land Use and Planning Data

3.3.1 Land use and planning data are key drivers for both the estimation of variable demand models, and their application to future scenarios. This section describes available data sources and, where relevant, their strengths and weaknesses.

Office for National Statistics Population Data

3.3.2 The Office for National Statistics (ONS) population data are published every year and provide estimates of UK population, by sex, age, and local authority district. The estimates are generated using a combination of complementary data sources, such as the previous census, household surveys, and administrative registers. The data are the official population estimates of the UK and are considered to be fully representative and unbiased.

Office for National Statistics Employment and Labour Market

3.3.3 The Office for National Statistics (ONS) provides data on employment and the labour market in Great Britain. The data provide information on employee working characteristics such as type of work, earnings, and workplace disputes. The data are largely sourced from the Labour Force Survey (LFS) and are released annually; many of the data sets are available as time series. Given the data are sourced from the LFS, there are expected to be statistical errors resulting from the survey sampling. The spatial granularity of the data varies between data sets.

Valuation Office Agency Floorspace Data

3.3.4 The Valuation Office Agency (VOA) provides data on the amount of floor space and number of properties by employment sector, property type, and rateable[2] value band. The data are provided by Lower Super Output Area (LSOA). The data are available from 2001 and cover only rateable properties in England and Wales, and missing floorspace data are infilled. Information about the data, including the imputing of missing data, is provided in the agency's webpage.

Department for Education School Census Data

3.3.5 The Department for Education (DfE) releases annual data on school pupil numbers and their characteristics by school and by local authority district within England. The data cover information such as class sizes, student ethnicities, and students with English as an additional language (EAL). The state school data are collected through a DfE school census. Data are also provided for independent schools, general hospital schools, and Alternative Provisions[3]. .

Driver and Vehicle Licensing Agency

3.3.6 The Driver and Vehicle Licensing Agency (DVLA) vehicle licensing statistics provide a complete record of all registered vehicles in the UK. The data contain a new record for each time a vehicle is registered in the UK. They include information about the vehicle such as make, model, and colour, as well as some information about the registered owner.

[2] Rateable properties are buildings or premises within buildings, appropriate for or used for single occupation. Where the rateable value is broadly the value at which a property may be let for one year

[3] Alternative Provisions are places that provide education arranged by local authorities for pupils who, because of exclusion, illness or other reasons, would not otherwise receive suitable education; education arranged by schools for pupils on a fixed period exclusion; and pupils being directed by schools to off-site provision to improve their behaviour.

3.3.7 The DVLA estimates that data are correct for 89% of the registered vehicles, and the remaining 11% have an error, almost exclusively due to registered owners incorrectly filling in vehicle registration forms. The data are provided at varying spatial levels, ranging from postcode district to regions across England (former Government Office Regions), and are released annually. A key limitation of the DVLA licence statistics is that the location is not necessarily where the vehicle is used or kept, it is based on place of registration.

Mosaic Geodemographic Data

3.3.8 Mosaic is Experian's system for geodemographic classification of households, based on 69 types within 15 groups, using data sources such as rurality, council tax band, and property value. Data are available on the percentage of people and households that fall into each group and type and are provided at LSOA level. Documentation about how the estimates are prepared is available on their website.

3.3.9 This data source has been included within the unit as it provides spatially disaggregate and detailed household demographic data. However, it is not clear how recent available data are or how frequently the data are updated. The practitioners should therefore consider and report the validity of this data if Mosaic is used.

Acorn Geodemographic Data

3.3.10 The Consolidated Analysis Centre, Inc. (CACI)'s Acorn data categorise postcodes into 6 categories, 18 groups, and 62 types, by analysing social factors and population behaviours to provide information on the types of residents in neighbourhoods. Within different geodemographic types, the dataset holds information on car ownership, as well as estimating the percentage of those who use different transport modes to travel to work. An example of government use is available on HM Land Registry's website and can provide further insights into strengths and weaknesses. Data should be used with care and ideally verified at some level using existing and additional evidence. Data for the current year are available by postcode for the UK and are updated regularly.

Index of Multiple Deprivation

3.3.11 The Department for Levelling Up, Housing & Communities (DLUHC) (previously known as the Ministry of Housing, Communities, and Local Government (MHCLG)) provides a ranking of all LSOAs in England in terms of deprivation, known as the Index of Multiple Deprivation. There are seven measures of deprivation used to derive the indices, including income deprivation, crime, and living environment deprivation. Data were first published in 1998, and updated releases have been produced roughly every four years.

AddressBase

3.3.12 AddressBase (premium or plus) is a dataset of residential and commercial property addresses within Great Britain. AddressBase premium provides information on Local Authority, Ordnance Survey, and Royal Mail addresses and their history. AddressBase plus provides more information than AddressBase premium about individual addresses (e.g. flats) rather than Royal Mail post points (e.g. houses divided into flats) and provides more information on property classifications.

3.3.13 The data are updated every six weeks and contain all addresses that have been granted planning permission. Data for Northern Ireland, the Channel Islands, and the Isle of Man are also available. A licence is required to access the data, although it is free upon registration for public sector customers.

3.4 Geographical and Network Data

Ordnance Survey MasterMap Topography Layer

3.4.1 The Ordnance Survey (OS) MasterMap Topography Layer provides a detailed GIS layer for Great Britain identifying land use features, including buildings with different purposes, services, waterways, and green space. The features captured are listed in a catalogue provided by the OS and include details down to individual buildings. Railway and tramway lines are available within the layer, this includes underground lines when they are close to the surface or open to the sky. The topography layer is updated every six weeks, with building heights updated every six months. Practitioners should review the data to identify any gaps and apply appropriate methods to address these if they exist. Users may be able to make use of the Public Sector Geospatial Agreement (PSGA) to access the data.

Ordnance Survey Terrain Mapping

3.4.2 The Ordnance Survey (OS) Terrain 50 and Terrain 5 datasets provide terrain models of Great Britain with contours, giving a surface model of the entire landscape, including major roads, large lakes, and estuaries. Terrain 50 provides a digital terrain model grid with a post spacing of 50m and is updated annually in July. Terrain 5 provides a spacing of 5m and is updated quarterly. Users may be able to make use of the Public Sector Geospatial Agreement (PSGA) to access the data.

Ordnance Survey MasterMap Highways Network

3.4.3 Ordnance Survey (OS) MasterMap Highways Network - Roads is a link-and-node network representing the road network of Great Britain. It classifies links, such as motorways and pedestrianised streets, as well as providing the Department for Transport's road classification. The Highways Network is updated approximately monthly and covers all roads in Great Britain together with ferry links.

3.4.4 The MasterMap Highways Network offers additional layers providing rich information about the network.

3.4.5 The MasterMap Paths Layer is a topologically structured link-and-node representation of the pedestrian path and ferry network in urban areas. It allows for connectivity to the road network included with the Roads and Routing and Asset Management Information layers.

3.4.6 Information on the function, name, and surface type of paths are available within the layer. Both the Roads and Paths layers combine Ordnance Survey geometry with authoritative data from the National Street Gazetteer (NSG), the Trunk Road Street Gazetteer (TRSG), and the Scottish Street Gazetteer (SSG). The practitioners should be aware that the data may include issues of inaccurate representation, errors, and omissions. The data should be checked, and any issues addressed as required.

3.4.7 The MasterMap Routing and Asset Management Information Layer contains the MasterMap base road network with information on the rights and restrictions along the network, advisory information such as the locations of tolls and level crossings, and asset management information such as physical characteristics and road maintenance responsibility.

3.4.8 Average speeds and speed limit information are also available from the MasterMap Highways Network with speed data. Average speeds are based on annual averages, therefore they should not be used to represent speeds for specific time of year. They are calculated using historical data based on telematics from over 135,000 vehicles (mainly cars, with some LGV and HGV). The average speed data are broken down across three daytime periods for weekdays, along with segregated weekend and evening periods. Users should be aware that road links with missing speed data are infilled using neighbouring link information.

Ordnance Survey Open Roads

3.4.9 Ordnance Survey Open Roads is a digital representation of Great Britain's roads. The links represent an approximate central alignment of the road carriageway. Attributes identify the roads that make up the Primary Route Network (PRN) and the Strategic Route Network (SRN), along with individual link lengths and information on the road type. The data include all roads except private roads and shorter cul-de-sacs and are updated twice a year in April and November. This data does not require a license and is free.

Office for National Statistics Boundary Data

3.4.10 The Office for National Statistics (ONS) provides boundary data files, describing boundaries and weighted centroid data for multiple boundary geographies in GIS format. Geographies include Output Areas (OA), Lower Super Output Areas (LSOAs), Middle Super Output Areas (MSOAs), Census, and electoral boundaries.

OpenStreetMap

3.4.11 OpenStreetMap (OSM) is a collaborative project to create an open source, editable map of the world. The geodata underlying the map are considered the primary output of the project and are fine-grained, including all streets and walkways. Structured data can be downloaded and queried for features including administrative boundaries, roads, points of interest, and land use categories. The data are crowdsourced, include details of updates and correction of features, and are updated daily. The practitioners should be aware that due to their crowdsourced nature, the data may include issues of accurate representation, errors, and omissions. The data should be checked and any issues addressed as required. Network data can be extracted from the OSM through the Openrouteservice APIs.

Major Roads Database

3.4.12 The Major Roads Database is a GIS database of major roads containing a series of links which are unique sections of road which make up the entirety of the major roads network in Great Britain. The database is maintained and updated by the Department for Transport using information from the Ordnance Survey MasterMap Highways Network. This database contains various fields of information about these links including link length, available traffic count information, and other physical characteristics for all links greater in length than 300 metres. The data covers major roads only and is unlikely to be comprehensive enough for most local models.

Openrouteservice Distance and Time Matrix API

3.4.13 Openrouteservice provide an open source API to generate a matrix of journey times and distances by road-based mode between coordinates defined by the users. Where possible, the coordinates should represent a point on a road, however, where they are not, a nearby point on a road is used to calculate journey times and distances. The journey times and distances are calculated from the underlying OpenStreetMap (OSM) data. The calculations of journey times and distances do not represent variation in time of day.

National Cycle Network

3.4.14 The National Cycle Network (NCN) is a network of more than 12,000 miles of signed routes for walking and cycling across the UK. The data provide information on route attributes, such as lighting and road quality, and are updated monthly. The data are primarily collected and updated by volunteers, and do not cover the whole UK road network, only routes deemed safe for cycling. Therefore, data should be checked for gaps and refined as necessary before they are used for cycle network development or similar purposes.

National Rail Timetable

3.4.15 The National Rail Timetable (NRT) is the timetable of all National Rail train services in Great Britain, owned and updated by Network Rail. The timetable is updated twice a year, in May and December, and hence there may be some instances where new stations or routes have opened but the associated timetable data are not included in the current release of the NRT. Historic timetables are not available, which needs consideration when building base year models that are more than a year in the past.

CIF Railway Timetable Data

3.4.16 The Common Interface File (CIF) is a means of transferring railway timetable data from Network Rail's Integrated Train Planning System (ITPS). The data are updated daily and cover all timetabled rail services across Great Britain for the current week. The data are available to download from the Rail Delivery Group (RDG) and the Association of Train Operating Companies (ATOC) by registered users.

Traveline National Dataset

3.4.17 The Traveline National Dataset (TNDS) is a collection of timetables for all bus, light rail, tram, and ferry services in Great Britain. The data are collected from a partnership of transport companies, local authorities, and passenger groups, and are updated weekly. Not all data are updated weekly, information is provided online about when each dataset was last updated. The data can contain duplicate services if the service crosses a regional boundary, and contain temporary route information.

Realtime Trains

3.4.18 Realtime Trains provides web-based train timetable information for the current day, historical train running data up to 7 days previously, and future trains for up to 90 days. The website displays working timetables, actual train performance information such as delays and cancellations, and short-term alterations to the timetable. Also included on the website are non-passenger services including empty stock and freight services. Data, such as availability of first class and catering, are available for some operators, with stock type and train length data available for most services.

3.4.19 The data are obtained from Network Rail Open Data Feeds combined with internal data feeds to the website and cover all train operating companies in Great Britain, and passenger, non-passenger, and freight trains.

Office of Rail and Road Station Usage Data

3.4.20 The Office of Rail and Road (ORR) station usage data provide estimates on the number of annual entries, exits, and interchanges at rail stations in Great Britain. The data are generated from rail ticket sales and hence there are derived estimates included for entries and exits to do with 'rover' tickets, season tickets, and multi-mode tickets. The data are released annually and historical data are available from 1997.

National Rail Enquiries Open Data Feed

3.4.21 The National Rail Enquiries (NRE) open data feeds provide real-time arrivals and departures of National Rail trains in Great Britain. They contain information such as real-time departure boards, and the Historic Service Performance (HSP) can provide this same information for the previous year.

Department for Transport Bus Open Data

3.4.22 The Department's Bus Open Data Service (BODS) provides timetables, locations, and fares for bus services in England. The data do not include Transport for London services. The data are collected from all public bus service operators following a change in the legal requirement in 2017 for operators to provide this information. The timetable and fare data are updated twice a day, and the detailed location of individual buses are updated every ten seconds. The data are available to download by bus operator and location from the Department for Transport. Historical data are available, although the length of time series available varies by timetable.

National Public Transport Access Nodes

3.4.23 The National Public Transport Access Nodes (NaPTAN) database provides data on all public transport access points in Great Britain. The data cover bus stops, rail stations, airports, ferry piers, and light rail stops. The data are updated regularly and provide public transport access points and their coordinates to allow GIS analysis.

3.4.24 The NaPTAN database system is a UK national standard sponsored by the Department for Transport and supports both the public registration of bus timetables by the Vehicle and Operator Services Agency (VOSA) and the data collection for the Transport Direct Portal.

Department for Transport Maritime Statistics

3.4.25 The Department for Transport's Maritime Statistics provide data on freight traffic to and from UK ports, carried on inland waters and around the coast of the UK, together with statistics on international and domestic passengers, estimates of the number of UK seafarers, and information on the number and types of shipping fleet in the UK. Data are collected from shipping lines, operators, shipping agents and ports.

3.4.26 Most categories cover all data in the UK, though this can vary. Port level statistics cover major ports in the UK which have cargo volumes of at least 2 million tonnes annually; ports in the Isle of Man and the Channel Islands are not covered. Sea passengers statistics cover ferry operators calling at UK ports, including cruise ships. Most data sources are released annually, with some updated quarterly. Historical data are available on request.

3.5 Tracking Data

Mobile Network Data

3.5.1 Mobile Network Data (MND) record the times, dates, and locations of anonymised devices interacting with cell masts. These data can be interpreted, aggregated, and expanded to represent travel patterns. The sample size and coverage of MND are large and capture day-to-day variability in travel patterns, assuming the data are collected over a sufficiently long period – the length of which requires negotiation with data providers. A rigorous process needs to be carried out to assign travel modes and purposes to the data, and there are limitations associated with fully capturing short trips; however, newer technology may allow the impacts of these issues to be reduced in the future. Practitioners need to work closely with MND providers to ensure robust data are collated for their specific purposes.[4] Extensive reporting on data validity is expected, building on guidance embedded in the Transport Systems Catapult's report for the Department for Transport.

3.5.2 A detailed description of the data, including strengths and limitations, and the purposes the data can be used for, is provided in TAG Unit M2.2 Base Year Demand Matrix Development. Other data sets, such as the Census for expansion, or GPS Tracking Data for spatial refinement, may be required to support the use of MND.

GPS Tracking Data

3.5.3 GPS tracking data are collected from personal GPS devices and those in vehicles. GPS data provide information on journey times, speeds, routeing, and can be processed to derive origin-destination (O/D) demand.

3.5.4 GPS data are sourced from users based on their decisions to use particular GPS enabled services, and this can cause bias towards the typical owners of the relevant devices and the type of trips made. In vehicle devices tend to be installed in newer or higher value vehicles or specific driver types, which are typically used for a greater number of trips and trips of longer distances, compared with the average vehicle on the road. The size of such biases is expected to decrease as more vehicles adopt the GPS devices, but the extent to which it is still present is unknown and must be checked before the data are used for modelling purposes. Commercial sensitivities and privacy laws limit the extent to which these biases can be determined.

[4] Project EDMOND provides an example application of MND, fused with complementary data sources to develop multi-modal demand matrices for a dense urban area. See the following link for details: https://aetransport.org/past-etc-papers/search-all-etc-conference-papers?abstractId=6136&state=b

3.5.5 Credible and robust methods are required to check and correct for the biases[5], for both speed and journey time, and particularly for OD matrix purposes. Speed and journey time data are typically less impacted by the potential biases, however, it is not recommended to use the data to develop car demand matrices or for speed and journey time data unless the data are shown to be suitable.

3.5.6 The sample of GPS tracking data is understood to comprise a high proportion of freight vehicles rather than the average vehicle composition, because freight vehicles are more likely to have GPS devices fitted. When using this type of data for highway modelling, practitioners may need to contact data providers to understand the underlying data that is used in the dataset. Practitioners may need to adjust the data to indicate the absolute speeds needed for highway modelling and appraisal purposes.

3.5.7 Modellers should provide evidence to justify suitability of the data for the intended use. For example, on congested urban roads where HGVs and cars are expected to have similar speeds it is more likely that the sample of vehicles from GPS tracking data is representative of the average speed, therefore use of the data may be justified. On the other hand, where there are differential speed limits for HGVs, or where acceleration rates are expected to affect average speeds on shorter road links, the higher proportion of freight vehicles in the sample is likely to result in average vehicle speeds being understated.

3.5.8 TomTom is a source of tracking data collected from fixed or non-fixed in-vehicle GPS devices used for navigation. The limited available evidence suggests that passenger cars are the main sample for TomTom data[6].

3.5.9 INRIX is a source of tracking data collected from in-vehicle GPS devices or mobile phones' Bluetooth (used as a tracking device). The data are split into three vehicle weight classes. Users should note that the vehicle type definitions are not directly consistent with definitions used by Department for Transport or the Driver and Vehicle Licensing Agency (DVLA) to distinguish light and heavy vehicle types. At the time of writing, The Department for Transport's Congestion Statistics team has a sample of INRIX data that can be shared for free with local governments, transport executives and crown bodies.

3.5.10 Teletrac Navman (formerly known as Trafficmaster) is a source of tracking data collected from in-vehicle GPS devices. It is understood that the GPS devices providing the data are mainly associated with fleet management and theft tracking. The journey origins and destinations are provided at LSOA level. A journey is defined as the time between ignition switching on and off. The data are collected from around 100,000 vehicles. A detailed description of the data, including strengths and limitations, and the purposes the O/D data can be used for, is provided in TAG Unit M2.2 Base Year Demand Matrix Development.

[5] An example of robust bias correction was used during the development of LGV trip matrices for the Second Generation of Regional Traffic Models (RTM2). See Tolouei, et. al., (2022) for details

[6] Fernandes, et. al., (2013)

3.5.11 Google's Distance Matrix API provides distance and journey time matrices, and routeing between points based on current and historical traffic conditions. The mode of travel can be specified based on an assumed mode in their model algorithm, with the option to limit transfers and walking distance when using transit modes (bus, rail, subway, train and tram). A current or future departure time can be specified, to determine the best route and journey time in expected traffic conditions, or a request can be made to use average conditions, although it is unclear how these are defined. The sample size used by Google is expected to vary by time of day and day of week, as is the representation of different modes in the dataset.

3.5.12 These data from Google provide vehicle journey time data, which are otherwise only available from a limited number of data sources, and are assumed to have a large sample size given the popularity of the Google Map's app. However, practitioners should be aware of the limited amount of information that is available on the data collection and processing steps when using the data source.

3.6 Ticket Data

Electronic Ticket Machine Data

3.6.1 Electronic ticket machine (ETM) data cover all public transport bus fares purchased electronically via tickets or smartcards. ETM data provide time of travel and can be obtained over long periods of time, thereby reflecting day-to-day variations. They can also relate to the network of services as a whole. However, ETM data will only provide trip records in terms of fare stages at which passengers board or alight, and fare stages may differ between different operators.

3.6.2 By definition the ETM data do not represent bus journeys completed without a ticket or ticket product. Ticket products include customers boarding with a pass or reusable ticket. In areas where the use of travel cards, concessions, and other pre-paid tickets are prevalent, ETM data may provide a less accurate picture of passenger movement, and corrections will need to be made to reflect any missing trips.

3.6.3 Smart ticketing has been introduced in many towns and cities throughout the UK in recent years. The data capture system transactions and can be processed to provide data records on the usage of public transport. Like ETM data, smart ticketing data sets may not represent all users of the system, for example if tap on or tap off is not required. The growing use of smart cards in public transport, and the ease with which this data can be accessed will make the processing and analysis of this data increasingly useful for understanding and representing travel patterns in modelling and appraisal.

3.6.4 ETM and smart ticketing data are held by transport operating companies so practitioners should contact public transport operators in their area of interest to enquire about the availability of the data.

Latest Earnings Networked Nationally Overnight

3.6.5 Latest Earnings Networked Nationally Overnight (LENNON) is a database of (almost all) rail ticket sales in Great Britain that is updated daily. It also acts as a clearing house by allocating the revenue from these sales across Train Operating Companies (TOCs) and non-TOCs. LENNON's principal purpose is therefore to act as a financial accounting tool. The data are managed by the Rail Delivery Group (RDG) and contain ticket sales data from all available industry ticketing systems. Currently, journey data from five TOCs are not included in LENNON but the associated revenue is included.

3.6.6 LENNON is principally concerned with revenue but also estimates journeys (and passenger miles). Therefore, as an input to public transport models, it is possible to develop rail passenger trip matrices from the LENNON database with caution and consideration of its limitations. Given that the data are not purely observed but part-synthesised, any matrices developed should be verified against independent data sources and adjusted as necessary.

3.6.7 LENNON provides earnings data by station-station flow, ticket type, and travel date. The LENNON journey data are derived (not observed) based on journey factor assumptions, which vary for each product or ticket type. In particular, assumptions are made about the number of journeys associated with season ticket sales, multi-modal tickets, and 'roamer' tickets. The derived journeys therefore need to be verified against independent data sources and adjusted if necessary[7].

3.6.8 Careful consideration is also needed when data are used at a more granular than annual level, because the dating of LENNON journeys (and revenue) does not necessarily correspond with date of travel. The transaction date, referred to as "date of settlement", is not necessarily the date of travel, but is usually the collection date for tickets on departure. Advance tickets, in particular, are settled at least one day in advance of the travel date.

3.6.9 In addition to the above, practitioners need to consider other limitations when using LENNON data in transport models, which are:

- the data do not cover tickets bought for travel on the major light rail or underground systems of the UK, including the London Underground, and hence these are infilled
- monthly and annual season ticket journeys and earnings (not sales) are spread across the period of the ticket
- the data only covers rail journeys and do not cover full journey details where a multi-modal ticket is bought (e.g. rail/bus/LRT travelcards)

[7] Steer and WSP (2019) found that season ticket journeys are overstated, with the level of overstatement varying by trip distance and level of non-season ticket competition

- the data only include the station-to-station journeys, not total end-to-end journeys
- the data contain two different definitions of derived journeys (i.e. operating journeys and passenger journeys) and hence underlying journey definition should be considered when journey statistics are interpreted
- the data take account of travel that takes place without a ticket (e.g. free concessionary travel, PTE rail only travelcards or fare evasion) but have limitations with flat fare concessionary travel
- the data are not segmented by purpose or car ownership
- the data are not collected in production / attraction (P/A) format
- the trip origin is assumed to be the point of sale of the ticket, and hence there may be definitional inconsistencies to do with direction of the journeys
- changes to retail channels (i.e. significant movement to digital) potentially affects the relationship between dates of settlement vs travel
- the relationship between Passenger and Operating Journeys is driven by the timetable, which has changed post-Covid pandemic therefore care is needed when comparing pre-Covid and post-Covid journey data

3.6.10 LENNON data are owned by the Association of Train Operating Companies (ATOC) and, although the Department for Transport has access to the LENNON database, the Department can only share these data with third parties working on its behalf. Any practitioners not conducting work on behalf of the Department and requiring access to LENNON data should first contact the ATOC through the Rail Delivery Group (RDG) at their website.

Rail Usage and Drivers Dataset

3.6.11 The Rail Usage and Drivers Dataset (RUDD) contains historic demand, revenue, and timetable information for around 25,000 station-to-station flows, at an annual level, accounting for 95% of rail revenue in Great Britain, from 1994 to 2019. It also contains socioeconomic / demographic variables which drive rail demand, mapped to stations. It uses data from sources including MOIRA, LENNON ticket sale data, and various government departments. The ticket data only include the station-to-station journeys, and not total end-to-end journeys. The trip origin is assumed to be the point of sale of the ticket, and hence there may be definitional inconsistencies relating to the direction of journeys.

3.6.12 RUDD is owned by the Department for Transport and managed by the Rail Delivery Group (RDG). Any practitioners not conducting work on behalf of the Department and requiring access to RUDD should first contact the Association of Train Operating Companies (ATOC) through the RDG at their website.

3.7 Data from Non-Intrusive Surveys

Department for Transport Rail Statistics

3.7.1 The Department's Rail Statistics provide summary data on capacities, demand, and performance for the rail network in Great Britain.

3.7.2 Rail Passenger Numbers and Crowding on Weekdays data cover aggregate passenger numbers by city and central London station, and by hour of the day, as well as Passengers in Excess of Capacity (PiXC) data for average peak periods which are also provided by train operator. The data are generally collected in Autumn of each year, as this is assumed to be when commuting demand is at its highest; this may not be true of all stations or services. The data are collected either through manual counts or train load weight systems.

3.7.3 PiXC is defined as the percentage of standard class passengers above the capacity of the train service into a major station. PiXC data are a measure of crowding in the AM peak into stations (07:00-09:59) and PM peak out of stations (16:00-18:59). The data cover crowding from 2011 onwards.

3.7.4 The data are based on a minimum of one manual count per service so sampling error needs to be accounted for where a small number of counts were taken – and given that the counts take place in Autumn, other corrections may be required to represent a typical or neutral day. The data are provided by major city (e.g. Birmingham and Manchester) so there is no information by station, apart from in London where they are broken down by major station (e.g. Liverpool Street and London Bridge).

3.7.5 Further rail data sources are emerging during the writing of this unit, and practitioners are advised to make further enquires to the Department's Rail Statistics unit, and to Network Rail.

Department for Transport Road Traffic Statistics

3.7.6 The Department for Transport's Road Traffic Statistics (RTS) provide summary traffic flow data by road type, vehicle type, and region in Great Britain. The data are collected from around 8,000 manual traffic counts and 300 automatic traffic counts each year. The manual counts are taken between 7am and 7pm on weekdays between March and October. Additional automatic traffic count data are used from National Highways, Transport Scotland, and Transport for London.

3.7.7 The RTS provide Annual Average Daily Flow (AADF) and traffic data for every junction-to-junction link on the 'A' road and motorway network in Great Britain. The user can search, view, and download these data for each count point on the major road network, and view regional summaries.

3.7.8 An interactive map on the website provides a mapped background to identify traffic flows in specific areas of the country. Data from a sample of points on the minor road network are also available to download from the website.

3.7.9 Complete data sets are also available from the website. These include road traffic counts, local authority traffic estimates, and traffic, speed, and congestion data. The data go back to 2000 and in some cases to 1993.

3.7.10 As the flow data is based on a number of types of observation, its statistical characteristics are unknown, but confidence intervals can be expected to be large. Care must be taken when using this data for model validation purposes.at individual site level.

Department for Transport Congestion Statistics

3.7.11 The Department's Congestion Statistics team publishes the Congestion Statistics datasets. These data provide information on average speeds, delays, and travel time measures across the Strategic Road Network (SRN) and local A roads in England. The primary use of this data is to report congestion levels and the reliability of travel times on the road network. These data can be used by transport modellers to validate their models.

3.7.12 These travel times on roads in England are derived from Global Positioning System (GPS) tracking datasets provided by Inseego (formerly known as CTrack) and INRIX. Access to these data can be given to practitioners working on behalf of the Department, National Highways, Crown Bodies, or any of its agencies by contacting the Congestion Statistics team.

Traffic Information System

3.7.13 National Highways' Traffic Information System (WebTRIS) provides data on journey times, speeds, and traffic flows on all major roads managed by National Highways in England. The data are collected through a combination of Automatic Number Plate Recognition (ANPR) cameras, in-vehicle GPS devices, and inductive loops in the road surface.

3.7.14 The data are collected in 15-minute periods and vehicle types are identified by vehicle length. Not all count sites have continuous data.

Trip Rate Information Computer System

3.7.15 The Trip Rate Information Computer System (TRICS) provides trip generation analysis in the UK and Ireland. The TRICS database includes traffic and multi-modal transport surveys, covering a wide range of development types. The database covers a wide range of separate count and mode types and includes comprehensive site details such as the site's local environment, size information, size composition and functions. TRICS is owned by a collection of local authority councils. The data incudes count of public transport passengers and pedestrians, but the primary data are traffic counts.

3.7.16 TRICS is mainly used to estimate the trip generation impacts of new developments, and the user must take care to filter similar developments that may be assumed to have a comparable distribution of arrivals and departures, by mode and during the whole day. The data are available for the individual count sites and require a licence to access.

Traffic Data from Video Analysis

3.7.17 Multiple sources provide data from survey cameras, where video data captured by the sensors are processed using machine learning technology to identify motorised vehicle types, as well as pedestrians and cyclists. Video camera analysis is also used in stations and within public transport vehicles. There is little known about how accurate the machine learning techniques are at classifying vehicles or tracking individuals, and as companies providing such services tend to protect their proprietary processes there is little information available on the quality of this data. Therefore, the data should be verified, such as through manual review of a sample of video recordings and be used with care. All verification and any corrections should be reported.

3.7.18 This is an emerging data source that is being increasingly researched and improved for use in a wide range of applications. The potential biases and processing errors are expected to be better understood and reduced with each application[8, 9].

Traffic Data from Local Authorities

3.7.19 Many local authorities have their own traffic and passenger counts. It is advised that users contact the authority in their area of concern to inquire about the availability and extent of traffic data.

3.8 Data from Interview Surveys

Household Interview Surveys and National Travel Survey

3.8.1 Household surveys provide the most complete picture of travel by residents of a study area, including walking and cycling. Outputs from these surveys can be segmented by the key variables of household type, person type, trip purpose, mode, and time period. However, building of trip matrices directly from household surveys is not generally practical on the grounds of cost - the method is inherently expensive per person trip recorded. The primary application of this type of data set is therefore in the segmentation of demand data collected from other sources, and for use in creating local car ownership and trip end models where the national models are not thought to be appropriate.

[8] See Gravett and Mundaca (2021)
[9] See O'Hare and Jackson (2020)

3.8.2 Household interview survey sample sizes are rarely sufficiently large to provide acceptably accurate estimates of trips between pairs of zones. However, household interview surveys are a rich data source, in the sense that activity-based behaviour (by all modes) can be linked to the characteristics of the household and travellers. This data source is useful, therefore, for demand model estimation. Given that respondents tend to rely on recalling previously made trips, there is a risk of response bias in conducting household surveys, for example underestimating short or irregular trips.

3.8.3 The National Travel Survey (NTS) is a continuous household survey that has been collected since 1988. It is a rich source of information on personal travel within England and provides information on travel behaviour over time. The survey is primarily designed to track long-term development of trends, both travel related and otherwise, and hence care should be taken when drawing conclusions from short-term changes or when using it at a spatially disaggregate level.

3.8.4 The NTS collects responses from all age groups (over 5 years old). Some population groups are excluded from the sample if they do not live in a private household, such as students in halls of residence and tourists – this must be accounted for when these groups make up a significant proportion of travellers. NTS data are collected via two main sources; interviews with people in their homes, and a travel diary that they keep for a week. Typically, the sample size every year since 2002 has been 8,000 households and 20,000 individuals, although this can vary by year.

3.8.5 Summaries of the annual survey results are available for free from the Department for Transport. Access to the full survey data records can be requested from the UK Data Service. The allowable level of detail in accessing NTS data depends on the project and end user of the data.

Census

3.8.6 The Census is a nationwide survey taken every ten years covering all households in England and Wales (separate censuses are taken for Scotland and Northern Ireland). The census data provide demographic and socio-economic information for the entire population, as well as the usual method of 'travel to work'. The data is assumed to be fully representative and unbiased. However, due to the phrasing of the question, the absolute number of 'travel to work' trips are overstated and are also biased towards those modes that are most-used, understating the share of modes that are used sporadically.

3.8.7 The 2021 census is unique in that it was collected during the Covid-19 pandemic, and the responses were requested to be the response at the time of collection. There are therefore likely to be additional inconsistencies than in a standard census due to unusual travel patterns. For example, university students were required to provide their current home location, which, for a significant number, was their parental home and not the usual term time address. Therefore, the results of the 2021 census should be used with extra caution and are unlikely to be relevant for long-term transport modelling without further corrections.

3.8.8 Advice on use of 'travel to work' data is included in TAG Unit M2.2 Base Year Demand Matrix Development. TAG Unit A4.2 Distributional Impact Appraisal provides more information on how the census data, and others, can be used for socio demographic information.

3.8.9 The Census is updated every 10 years and at the time of collection is considered to be the most complete representation of the population and their characteristics. However, given the frequency with which the data are updated, consideration should be given to the validity of the data for the relevant modelling years and its suitability should be reported.

Civil Aviation Authority: Survey of Airport Passengers

3.8.10 The UK Civil Aviation Authority (CAA) undertakes annual surveys of airport passengers at a selection of UK airports. The airports generally cover the largest UK airports, with smaller airports surveyed less frequently. The passengers are asked about their personal characteristics, such as age, income level, and home location, as well as information about their trip, such as mode of transport used to access the airport, trip origin and destination, and group size.

3.8.11 Summaries of the annual survey results are available for free from the CAA. Access to the raw data can also be provided for a fee, depending on the number of records requested. Sample size should be considered, and raw data should only be processed and used at a level of detail that satisfies a minimum sample size requirement appropriate for the intended use(s).

3.8.12 The CAA also maintains punctuality statistics which are logs of total scheduled flights, departure and arrival times, and flight cancellations for all airports around the UK. These annual and monthly statistics are available to download for free.

International Passenger Survey

3.8.13 The International Passenger Survey (IPS) is a continuous survey (conducted on 362 days a year) at major ports of entry to, or exit from the UK. Voluntary face-to-face interviews with a random sample of passengers are carried out to collect information about the passengers' travel purpose, place of residence, and other trip characteristics. This data is used to produce estimates of visits to the UK by visitors from overseas, visits overseas by UK residents, and the length of time spent on visits.

3.8.14 The interview data are expanded to represent the actual number of passengers at the ports within the reporting period, and hence practitioners should take care when the data was based on a small sample. The data can be accessed from the UK Data Service for non-commercial purposes.

Continuing Survey of Road Goods Transport

3.8.15 The Continuing Survey of Road Goods Transport (CSRGT) records trip information made by HGVs registered in Great Britain, segmented by vehicle type and commodity group, including empty trips. The survey is conducted on around 230 vehicles per week, the results of which are expanded to the Driver and Vehicle Licensing Agency (DVLA) HGV population. The results of the survey generally tend to understate HGV trips but provide useful information on trip length and distribution patterns at a spatially aggregate level.

National Travel Attitudes Survey

3.8.16 The National Travel Attitudes Survey (NTAS), previously known as the British Social Attitudes Survey, is a Department for Transport survey collecting information on public attitudes to travel and transport in Great Britain covering subjects including road safety, barriers to cycling, and travel during the Covid-19 pandemic. The first data, from respondents of the National Travel Survey (NTS) aged 16 or over who consented to further contact, were collected in 2019. The results are provided by regions across England (former Government Office Regions).

3.8.17 The recent surveys suggest that the sample size is around 2,500 respondents, hence the practitioners should consider sample size before disaggregating results to a fine level. The data may overrepresent individuals with more spare time available, and who hence may consent to further contact from the NTS.

National Highways and Transport Network (NHT) Public Satisfaction Survey

3.8.18 The NHT Public Satisfaction Survey (PSS) collects public perspectives on, and satisfaction with, highway and transport services in Local Authority areas. Conducted annually since 2008, the survey also collects data on the use of different transport modes and access to services. Questions on topics such as road safety, the condition of highways, street lighting, and cycle routes are included. The survey in 2023, collected data from 410,000 households across Great Britain across 111 local authorities.

3.8.19 The PSS has been collected annually since 2008. Not all local authorities have data for all years of the survey. The practitioners should consider sample size before using any results at a disaggregate level.

National Rail Passenger Survey

3.8.20 The National Rail Passenger Survey (NRPS) was conducted twice a year by Transport Focus from 1999 until 2020, with the most recent sample covering around 19,000 respondents. It collected passenger opinions on their satisfaction with rail travel and was taken from a representative sample of journeys across the Great British rail network. Survey data were collected by paper and online surveys, with weighting applied according to response rates from different train operating companies, as well as journey purpose and time of week (weekday or weekend). The NRPS was not carried out between 2020 and 2022 due to the reduction of people using rail as a mode (because of the 2020 pandemic). However, Transport Focus has launched the Rail User Survey in April 2022, with a different set of questions and apparently less spatial detail.

3.8.21 The NRPS survey responses refer to the origin-destination (O/D) journey of the respondent. The sample size is small when looking at station or O/D level data, and the data should be grouped to a more spatially aggregate level to provide useful insights. The practitioners should understand the sample size of the responses before considering using data at a station or flow-specific level. In most cases, a high level of aggregation is required.

Department for Transport Bus Statistics

3.8.22 The Department's bus statistics provide data on trends in bus demand, mileage, and fares across Great Britain. The data are collected from an annual survey of over 500 bus operators, as well as Transport for London (TfL). The data are released annually or quarterly and are available to download.

3.8.23 The data are aggregate. They provide average or total statistics by year and by a variety of geographies, such as local authority, with fare indices also provided by month. There is no information on time of day, route, or operator.

Bus Passenger Survey

3.8.24 The Bus Passenger Survey was an annual survey of bus passengers conducted by Transport Focus, and was halted in Autumn 2019. The survey measured passenger satisfaction with their local bus service, on aspects such as the bus stop, waiting for the bus, and value for money. The survey was collected via an in-person form or online, with an approximate survey size of 50,000 passengers.

3.8.25 The survey only collected information from passengers aged 16 and over, and within England (excluding London). Data were collected during three weekday time periods (AM, PM and off-peak) and during the weekend. Methodological issues are described in a report.

Strategic Roads User Survey

3.8.26 The Strategic Road Users Survey is an annual survey from Transport Focus for those driving on England's motorways, and major 'A' roads. The survey collects data from around 8,000 individuals on user satisfaction on a range of aspects, such as road surface, journey time and safety. Households are invited to complete the survey, to form a representative sample of the driving population within England. The collection method was changed from face-to-face to online due to the Covid-19 pandemic.

Living Costs and Food Survey

3.8.27 The Office for National Statistics (ONS) Living Costs and Food Survey collects data on spending patterns and the cost of living that reflect household budgets. The survey is conducted annually throughout the year with a sample size of 5,000 to 6,500 households. Data are collected via face-to-face household interviews as well as respondents being asked to keep a diary of daily expenditure for two weeks. Those invited to the survey are a sample of the over16 private household population, excluding the Scottish offshore islands and Isles of Scilly, and results are weighted to adjust for non-response and to gross to population estimates. The selection of postcodes chosen for the survey aims to limit bias within the data, but ineligible addresses or households reduce the response rate and may introduce bias in the dataset.

Time Use Survey

3.8.28 The UK Time Use Surveys were conducted in 2000-2001 and 2014-2015, by the Office for National Statistics (ONS) and the National Centre for Social Research (NatCen), respectively. The main aim of the survey was to measure the amount of time spent by the population on various activities, though demographic, economic and social information about households and individuals were also collected.

3.8.29 Data are collected via a time diary, usually over a single weekday and weekend day. The sample is based on households, and household members aged eight years and over. Weekly work schedules are also collected from those respondents in paid work. In the 2014/15 survey a total of 9,388 individuals in 4,238 households provided 16,553 diary days. Weighting was applied to account for effects arising from the survey design.

3.8.30 There is also time use data that has been collected in the ONS Time Use Survey (OTUS) on an experimental basis since March 2020. Practitioners should take care when using the data to understand changes in the methodology over time.

Annual Survey of Hours and Earnings

3.8.31 The Annual Survey of Hours and Earnings (ASHE) (previously New Earnings Survey) provides data on the level and distribution of earnings and worked hours for jobs within the UK. The data are based on a 1% sample of jobs from the HM Revenue and Customs (HMRC) Pay as You Earn (PAYE) records and are available by industry, age group, and local authority.

3.8.32 The data do not cover jobs not on the PAYE scheme, individuals that are self-employed, or employees that are not paid during the sample period. No method is used to infill missing responses or exclude outliers, and practitioners should ensure that, when using this data, appropriate corrections are made. Aggregate data are available from the Department for Transport.

Business Register and Employment Survey

3.8.33 The Business Register and Employment Survey (BRES) is an annual employer survey on the number of jobs held by employees by employment industry. The sample size of the survey is approximately 80,000 businesses. BRES data cover a subset of employers across the UK. The data are provided at Lower Super Output Area (LSOA) level or more aggregate levels depending on the access licence.

3.8.34 The BRES data cover trading businesses within the UK with employees who are registered for VAT or Pay-As-You-Earn (PAYE). Within the survey part-time employment is considered as 30 hours or less of work a week. Summary statistics are available online, with more detailed data subject to a licence agreement. The data do not include businesses which are not registered for VAT or PAYE, with variation between years of the survey as to what is included in the questionnaire. Practitioners should recognise that around half of UK businesses trade without being registered for VAT or PAYE, although their number of employees is much smaller.

Roadside Interview Surveys

3.8.35 Roadside interview (RSI) surveys have traditionally been used as a source of data to build highway demand matrices. For a discussion of their strengths and limitations, and an overview of the process of developing matrices from RSI surveys, please refer to TAG Unit M2.2 - Base Year Demand Matrix Development. For general advice on conducting highway surveys, including RSI surveys please refer to Section 4.3.

4. Collecting New Data

4.1 Introduction

4.1.1 This section provides specific advice on conducting different types of bespoke transport surveys to collect data for individual studies. The advice in this section should be read in conjunction with the general advice on data collection in Section 2.4.

4.1.2 The advice in this unit does not cover all possible survey types or methods. As new data sources and collection technologies emerge, survey methods that are not covered here may become available. If the survey method is suited to the modelling requirement, data collection techniques should not be limited to those discussed in this unit. Other data sources should be subject to the same level of scrutiny and reporting.

4.2 Demand Behaviour Surveys

Revealed Preference and Stated Preference Surveys

4.2.1 Revealed Preference (RP) refers to observations of actual behaviour, for example the mode choices that travellers currently make or made in the past. Stated Preference (SP) refers to observations of hypothetical behaviour under controlled experimental conditions.

4.2.2 RP data are inherently more credible than SP data and their use, if only partially, will strengthen the credibility of demand forecasts in the appraisal framework. A scheme that introduces a new alternative, for example a new mode or trip attractor, would imply a need for SP analysis, since RP data would be by definition unavailable for such a context. Developing a bespoke choice model, therefore, often requires SP surveys and analysis, which is generally strengthened by controlling the SP responses to observed and more voluminous RP data (when it is or becomes available).

4.2.3 RP data can be obtained from SP respondents, from postcard surveys (an under-used and relatively inexpensive approach), from home or phone interviews, travel diaries, as well as from the National Travel Survey and Census. Sometimes an RP survey can be used to also elicit SP responses. Increasingly, passively collected data such as Mobile Network Data, is used to control SP responses.

4.2.4 The collection of RP data is not without problems. There are often biases in respondents' self-reported data, underestimating the costs of their travel choice and overestimating the costs of their alternatives. To overcome these problems, it is sometimes necessary to use explanatory variables from network models and published timetable data rather than rely on respondents' perceptions. Even where respondents' reported data are modelled, there is often a considerable amount of missing data (mainly about alternatives) which needs to be collated separately.

4.2.5 For more information on SP and RP surveys see also Supplementary Guidance - Bespoke Mode Choice Models.

4.3 Highway Surveys

4.3.1 A large amount of highway-related data is available from national data sources, as described in section 3. For many applications, additional data collection is desirable, for example to update out-of-date information, to close gaps in screenlines and cordons, and to provide local context.

Traffic Count Surveys

4.3.2 Manual Classified Counts (MCC) are required to break down traffic flows by vehicle type. This information is particularly important in an urban area, where the mixture of vehicle types may vary significantly by direction as well as at different times of day. Despite the term manual, increasingly cameras are used to collect this information, although the associated video analysis could be carried out by humans or through machine learning. See paragraph 3.7.17 for further notes on the latter.

4.3.3 Turning counts (the measurement of flow from each entry point to each exit point) at road junctions are required for the accurate modelling of junction movements in a congested assignment model. In urban areas, there will often be a need to collect more turning count data than for an inter-urban model, because there is a likely to be greater opportunities for re-routeing in an urban area, and hence, the greater need for detailed knowledge of observed conditions.

4.3.4 Turning counts are carried out in the same manner as manual classified counts on links, except that more enumerators or cameras are generally required. They must cover the whole of each peak period but need only cover representative parts of other time periods, depending on the time periods being modelled. Where an inter-peak model is representing an average inter-peak hour, a 12-hour period covering the two peaks and the inter-peak would be required. The vehicle classification used may be simpler than the one used for link surveys, provided that it is again compatible with the model classifications. For more complex or larger junctions, video or ANPR surveying methods may need to be employed.

4.3.5 Automatic Traffic Counts (ATCs) are used for all the purposes mentioned in paragraphs 4.3.1 to 4.3.4 and, in addition, can be used to monitor traffic flows and to provide information about the relationship between survey day traffic and more stable longer term flow levels that would be more suitable for modelling and appraisal. Automatic counts can also be used to provide information about local 12-, 16-, 18- and 24-hour flow ratios, and daily and seasonal traffic variations, all of which are required to estimate average daily traffic flows from shorter period data.

4.3.6 Little definitive work has been published concerning the accuracy of traffic counts by automatic traffic counters. Experience suggests that the errors are machine and (particularly) installation dependent. TRRL Supplementary Report SR 514 "Estimation of annual traffic flow from short period traffic counts" contains some useful results on the efficiency and accuracy of annual estimates from short period counts. For longer term counts, the frequency and diligence of the station monitoring and servicing will be crucial.

4.3.7 Good quality counts are required for validation (as well as for calibration). The notion of using good quality counts for calibration and poorer quality counts for validation, or vice versa, should not be considered. This is also an important consideration when mixing newly collected ATCs, for which confidence intervals can be calculated, with data from national sources described in section 3, for which this information is generally not available.

4.3.8 Ideally, 95% confidence intervals for traffic counts should be calculated directly from the collected data. In the absence of this information, the following 95% confidence intervals should be assumed, for surveys that have been designed well and carried out correctly:

- Automatic Traffic Counts: total vehicles: ± 5%
- Manual Classified Counts: total vehicles: ± 10%

  Cars and taxis: ± 10%

  Light goods vehicles: ± 24%

  Other goods vehicles: ± 28%

  All goods vehicles: ± 18%

4.3.9 The ATC confidence intervals relate to counters with tube vehicle detectors. Counters with inductive loops or cameras may achieve greater levels of accuracy; the latter's accuracy depends on how the images are analysed, and practitioners should investigate if providers have evidence for the accuracy achieved. The accuracy of radar counters is less certain but may be assumed to be the same as that of tube counters.

4.3.10 Splits between light and heavy vehicles obtained from ATCs on the basis of a 5.2m vehicle length have been shown to be subject to wide margins of error and should not be relied upon. The National Highways WebTRIS database is based on a 6.6m split, that is deemed more appropriate. Vehicle classification should aim at the best trade-off between the number of classes and the reliability of distinguishing between them with confidence. Providers should be able to advise. Where possible spot checks against MCCs should be undertaken to verify vehicle splits derived from ATCs or WebTRIS.

4.3.11 It is normal practice for MCCs to be carried out on a single day but ATCs should be conducted for at least two full weeks. ATCs carried out for two-weeks or longer will capture, to an extent, the day to day variability. The confidence intervals for counts will be narrower than those listed above if more observations are carried out (for example an MCC carried out on two days) but will be wider for periods shorter than one day (for example, the morning peak). Extended ATC periods will allow the identification of 'outliers' for other count data and assist in the identification of typical or neutral conditions, and the calculation of annualisation factors.

4.3.12 Turning movement counts at junctions are normally single day MCCs (or video surveys / ANPR at more complex junctions). Their usefulness is increased by supplementing them with the two-week ATCs taken for either individual turns or on the junction entry and exit arms.

4.3.13 All data should be checked to identify and remove any that might have been affected by unusual events. Where data quality is suspect, the data should be investigated thoroughly and, if necessary, rejected.

Factoring Traffic Data

4.3.14 This section deals with the factoring of traffic count data from one modelling period to another. Every factor has an associated reliability, and although the aim of factoring is to increase the usefulness of data, the downside is that factoring increases the confidence interval of the result. Factoring should therefore be kept to a minimum and the factor with the lowest coefficient of variation should always be chosen where a choice of factors is available.

4.3.15 National factors are derived from databases which are usually larger than those a local study can generate. Whilst it is possible to derive factors locally, a full understanding of the accuracy of such factors is desirable to ensure that local conditions are indeed significantly different from the national average. This will involve not only an estimate of the coefficient of variation of the locally derived factors, but also an examination to determine whether the local factor is significantly different from the national one.

4.3.16 On some occasions more than one factor will have to be cumulatively applied in stages. TAG Unit M2.2 - Base Year Demand Matrix Development provides advice on such factoring and how the resulting reliability of the combined results is calculated.

4.3.17 There are occasions when practitioners need to obtain an estimate of (usually) AADT from a short period count (normally 12 hours). In such cases the recommended approach is to derive conversion factors from nearby locations where AADT data are available.

Journey Time Surveys

4.3.18 Journey time measurements are an essential part of assignment model calibration and validation for most urban traffic appraisals since the majority of scheme benefits tend to be related to journey time savings. Comparison of observed and modelled journey times gives a measure of the appropriateness of the speed-flow relationships for a capacity restrained assignment, as well as the junction delay calculations for a congested assignment model that includes junction representation. Journey time surveys may also be used to identify junctions which exhibit high levels of delay and that need to be modelled in detail.

4.3.19 Moving Car Observer (MCO) surveys may be undertaken, but these are expensive and require significant resource, data may be sourced from tracked vehicle data (using GPS devices), and between ANPR cameras. MCO observations record median speeds but are likely to understate average speeds where traffic exceeds speed limits and may require many observations where journey times are particularly variable. Vehicle tracking data may be from a small sample of vehicles and potential bias should be considered and addressed. It is often appropriate to use the median rather than mean measure of average which is less sensitive to exceptional behaviour of a few users or of road conditions. The data should be demonstrated to be of sufficient accuracy (see paragraph 4.3.14) and not materially biased. In the validation step, comparisons with modelled travel times will show how well total link times are modelled.

4.3.20 For general purpose models, the routes for the validation of journey times should cover as wide a range of route types as possible and cover the fully modelled area in the model as evenly as possible. For models developed for the appraisal of specific interventions, routes should include those on which it is expected traffic will be affected by the scheme, as well as covering the route including the proposed scheme itself, if appropriate.

4.3.21 The validation routes should be neither excessively long (greater than 15 km) nor excessively short (less than 3 km). Routes should not take longer to travel than the modelled time periods (although, a few minutes longer is unlikely to be problematic). Start times should be staggered, particularly if MCO runs are undertaken on the same day. For models of actual peak hours, journey time routes ought to be no longer than about 40 minutes to allow some staggering of start times.

4.3.22 During the survey, the total travel time should be recorded separately for each road section between major junctions and, because junction delays form an important part of travel time, a separate note should be made of the delay time at each junction. Ideally, delay should be assumed to start once instantaneous speed falls below a chosen speed, say 15 kph (10 mph). Journey time runs, in both directions and in each model time period, should be made over a period of several days. Variations in travel times during peak periods should be taken into account by staggering start times, to represent fairly conditions over the time period as a whole.

4.3.23 In the case of journey times for all vehicles combined, sufficient MCO runs should be undertaken so that the 95% confidence level of the mean of the observations is ± 10% or less over a route as a whole. Four initial journey time runs (preferably each on a different day) should be made for each route, direction and model time period, and the results used to assess the variability of journey times in each case. Further runs must be made for those routes, directions or time periods in which the variability falls outside the acceptable range. In urban areas, where journey times can sometimes be erratic, this may lead to a large number of runs being required. If a satisfactory level of consistency has not been achieved after 12 runs, the results should be accepted, and a special note made in the survey documentation. The accuracy of journey time data derived from other sources, such as tracked vehicle data or WebTRIS, should also be determined and taken into account when making comparisons between modelled and observed times.

4.3.24 When using separate speed/flow relationships for light and heavy vehicles, journey times applicable to these two classes of vehicle would be desirable, but not essential. For light vehicles, appropriate journey times may be obtained by restricting observations to light vehicles. For heavy vehicles, other techniques, such as video registration plate surveys may be required.

Roadside Interview Surveys

4.3.25 The roadside interview (RSI) survey process is very disruptive, which requires careful planning and management of risks. Safety to the staff involved and the general travelling public must be mitigated, often in conjunction with the police. Practitioners should note that RSI surveys and their associated mitigation measures can be very expensive and require experience and planning.

4.3.26 Practitioners should note that roadside interviews cannot be conducted on motorways. In addition, it is often impractical to conduct roadside interviews on some all-purpose roads, especially where they carry high levels of traffic. Relocation of interview sites may address these problems, for example, interviewing on motorway entry slip roads may provide an alternative to interviewing on motorways themselves. All sites require police permission and supervision, with site safety and prevailing traffic speeds a prime consideration for approval. Many sites will require temporary traffic orders to suspend parking, or bus lanes, or to impose temporary speed limits. Practitioners should be aware that these require notice periods and action by the relevant highway authority.

4.3.27 RSI surveys should always be associated with a MCC survey carried out on the day of the RSI survey and a minimum two-week ATC survey. Roadside interview records should be expanded by vehicle type from the MCC on the day of survey. The surveys should be scheduled to minimise potential for diversions to avoid the interview site, thereby making the MCC on the day of survey atypical.

4.3.28 Classified counts are required at every roadside interview site (where undertaken to understand demand patterns), and on parallel roads that are likely to be used as diversion routes during the RSI surveys, to expand the interview sample to the total traffic flow in the corridor as a whole. Counts should be carried out in both directions on the survey day, even if interviewing is only in one direction, and should extend over all model periods. If automatic counts indicate that traffic flows at a roadside interview site were influenced by the presence of the interview survey, further manual classified counts should be made on a different day. If necessary, these alternative counts can then be used to expand the interview data to a more representative traffic flow. The vehicle classification used should correspond with that used in the interview survey itself, and this in turn should be compatible with the vehicle types represented in the traffic model.

4.3.29 Addresses in the processed RSI data should be coded either to Ordnance Survey Grid References (at least six digits) or to postcodes. Adopting these minimal standards will ensure that data can be re-used by others at a later date, who may want to use the data in a different zoning system.

4.3.30 Guidance is available on calculating required sample sizes when conducting RSI surveys. TAG Unit M2.2 Base Year Demand Matrix Development contains advice regarding the question on estimating the sample size needed to give results to the level of accuracy needed.

4.4 Public Transport Movement Surveys

4.4.1 This section provides advice on conducting public transport movement surveys. The advantages and disadvantages of each survey type are also addressed.

4.4.2 The information required about passenger movements for an assignment model is:

- (true) origin and destination and preferably also first and last station/stop
- access mode to public transport, including any costs
- time of travel
- type of ticket

4.4.3 Other information, such as trip purpose and car availability for the journey, will usually be required for the development of the demand model. While other sources, such as a household interview survey, may also provide this information, it will normally be cost-effective to collect this additional information in passenger movement surveys when the data can be linked to an actual journey made.

4.4.4 The main sources of information about passenger movements are as follows:

- interviews with passengers, which may be by means of face-to-face interviews with passengers, either on-board the public transport vehicles or as they wait to board at stops and stations, or by means of self-completion questionnaires
- electronic ticket machine and smartcard data

4.4.5 The relative merits of these sources in providing the required information are as follows.

Face-to-Face On-Board Surveys

4.4.6 Face-to-face interviews with passengers on board, providing the sample size is adequate, can provide good quality data. The sample should be selected by the interviewer rather than being self-selected. This method will enable origin and destination addresses to be provided, along with all the other information that may be required, such as access mode (and trip purpose and car availability, if required). A sample of interviews should be obtained on each service in the modelled area, thereby enabling a complete picture of travel (once the sample data have been expanded to passenger counts).

4.4.7 In face-to-face interviews, the data can be collected from most individuals, including those with a difficulty understanding a questionnaire, and juvenile travellers (who can only be interviewed in the presence of parents or guardians), which is not the case with self-completion surveys. In addition, non-response biases are reduced compared with self-completion surveys. The method also allows the surveyor to probe for particular information so that exact details of origin and destination can be discovered.

4.4.8 Depending on passenger flows, passenger counts can be combined with on-board surveys. Often, on board face-to-face surveyors will be able to count the numbers of boarders and alighters at each stop before selecting respondents for interview. However, they will not be able to do so on crowded vehicles nor where each individual interview is likely to over-run the time between stops, therefore, separate surveyors to conduct counts are usually required especially during peak periods.

4.4.9 For multiple door vehicles such as trains, there may be substantial differences in passenger numbers counted at different sets of doors. Doors near station entrances will have a higher number of boarders than more remote doors. Because of this, an accurate level of count detail will only be available if counts are conducted on each set of doors, on-board or at stations.

4.4.10 In survey design, both a minimum absolute number of travellers, and a minimum proportion of all travellers should be interviewed. Usually more passengers will board at each stop during peak hours than inter-peak, and therefore surveyors will each be able to interview a smaller proportion of travellers during busy periods compared with inter-peak hours. This imbalance can be addressed by conducting an increased number of survey shifts during peak hours and by weighting data according to passenger counts.

4.4.11 One of the biases that can arise from face-to-face surveys relates to length of journey - the longer the journey, the more likely that a passenger is to be sampled. To counter this bias, those travellers who are making only short hop journeys should have an equal chance of being sampled as those making longer journeys. To do this, the surveyor should select respondents by choosing a random passenger from those boarding at each stop, platform or station.

4.4.12 Consideration must be given to the choice of public transport routes to be sampled. Ideally, each public transport route serving the modelled area should be surveyed at least once. These should be surveyed over the whole day period for which the model will apply. Information on bus, light rail and tram services can be found using the Traveline National Dataset (TNDS) referred to in paragraph 3.4.17.

4.4.13 This survey method can only be conducted with the permission of the bus or train operating companies.

Face-to-Face At-Stop or At-Station Surveys

4.4.14 Face-to-face interviews with passengers at stops and stations can provide the same scope and quality of information as on-board interviews. However, interviews are required at all stops and stations for a complete picture of travel in the modelled area. This is often uneconomic and so surveys of this kind are best targeted at major generators or attractors of travellers which, in practice, means town and city centres. It is important that all stops and stations in the targeted areas are included in the survey. If the operators will not allow surveys to be conducted on board their vehicles, the modeller will have no option but to rely on a combination of at-stop or at-station surveys and electronic ticket machine data.

4.4.15 While interviews can, in principle, be conducted with boarders and alighters at bus and light rail stops or train stations, people alighting are likely to be much less willing to be interviewed than those waiting for a service to arrive.

4.4.16 As with on-board face-to-face surveys, greater proportions of inter-peak travellers may be interviewed than peak travellers. The important consideration is that the sample size during both the peak and inter-peak periods should be sufficiently large to support the creation of acceptably reliable trip matrices for these periods.

4.4.17 Unlike on-board surveys, issues of long and short journey biases do not apply in that those making short hops are as likely to be approached for interview as those making longer journeys.

4.4.18 However, the data collected by this method are likely to be biased in that those passengers that arrive at the stop or station immediately before the vehicle arrives ('runners') will be under-represented. Where there are infrequent regular services, such as on trains, then this can lead to biases between regular and irregular passengers. Regular users of the service may time their arrival to the station more closely to the departure time. This bias may not be as prevalent on regular services where departure times cannot be easily predicted, such as for most urban bus routes.

4.4.19 As with on-board surveys, data collection should be conducted for the entire period for which the transport model will apply.

4.4.20 Whereas permission is often not required for surveys at bus stops, it will be required for platform or station entry or exit surveys.

Self-Completion On-Board Surveys

4.4.21 Self-completion surveys should be regarded as a means of last resort. While, in principle, the scope of the information which may be gathered by this means is as comprehensive as that obtained by face-to-face interview, the quality is likely to be poorer in a number of respects. First, some information may not be provided. Secondly, some information may be inaccurate if the questionnaire is completed some time after the journey was made. And, thirdly, the sample of respondents will be self-selected and may therefore not be adequately representative of all travellers.

4.4.22 The method involves the distribution of self-completion questionnaires to passengers as they board vehicles. Often these questionnaires would be collected from passengers as they alight from the vehicle.

4.4.23 Although forms should be distributed with reply-paid envelopes, in order to maximise response rates, it is recommended that measures be taken to ensure that forms are completed during the journey and returned to the surveyors when alighting. For this reason survey forms should be very brief and uncomplicated. Pens or pencils should be distributed to boarders along with the survey forms.

4.4.24 The main advantage of the self-completion survey compared with face-to-face interviewing is that all passengers have an equal likelihood of being given a survey form, regardless of whether they are making long or short journeys and whether travelling during peak or inter-peak times. However, return rates do vary due to self-selection biases.

4.4.25 Where self-completion surveys are collected from, as well as distributed to, passengers, high rates of return can potentially be achieved. However, large proportions of these forms will be incomplete. Response rates to individual questions in bus surveys can vary from between 40% and 70%. Response rates to questions may be higher on trains.

4.4.26 The same issues regarding which routes and times of day to survey apply to the self-completion on-board survey as to the face-to-face interview on-board survey. Similarly, permission is necessary from operating companies for surveyors to work on board vehicles.

Self-Completion At-Stop or At-Station Surveys

4.4.27 This method is conducted by distributing self-completion forms to passengers as they enter or exit stations or as they wait at or alight at bus stops.

4.4.28 The main advantage of this method compared with an at-stop face-to-face interview is that more passengers are likely to accept survey forms than would be willing to participate in an interview. This will apply to both those who arrive at stops or stations just before the train/bus arrives and also to those on time-critical journeys. Bus passengers waiting at stops would be more likely to accept a survey form even if their bus is approaching or already at the bus stop.

4.4.29 As with on-board surveys, at stop and at station surveys will suffer from inaccuracies and limitations to the data that can be collected. Depending on the survey materials, reported return rates of distributed questionnaires have been as low as 15% to 25%. Furthermore, some questions, such as on income, will often not be answered.

4.4.30 Again, sampling biases may occur depending on which stations and stops are surveyed and what times of day are covered. Litter generation of discarded survey forms can be a significant problem. Permission must be sought from station managers to distribute survey forms within stations.

Summary of the Preferred Approach

4.4.31 Table 1 sets out the main advantages of different public transport survey methods, along with biases that might occur, the practical difficulties of the method and likely response rates. Details of when each different type of method is most appropriate are also shown.

4.4.32 In summary, the best approach is for an adequate sample of face-to-face interviews to be conducted on-board a sample of public transport vehicles on each service in the modelled area. However, operators may not allow interviews to be conducted on their vehicles.

4.4.33 The next best approach is to conduct face-to-face interviews at stops and stations. While comprehensive coverage of railway stations may be affordable, it is likely that, to be cost-effective, surveys at bus stops would have to be confined to particular areas. This means that some other source of data will be required to supplement the at-stop surveys, such as passenger counts, electronic ticket data or mobile network data. With all methods, it is important that the survey is carefully managed to ensure public and surveyor safety, while maintaining the quality and reliability of the data being collected.

4.4.34 TAG Unit M2.2 Base Year Demand Matrix Development provides advice on the combination of survey data together with potentially complementary Electronic Ticket Machine (ETM) and smart ticketing data. Not all operators are willing to make their information available, especially in circumstances where the proposed scheme would compete with the operator's existing services. In these instances, self-completion questionnaires issued at stops may be used in combination with face-to-face interview at-stop surveys. In such an approach, face-to-face interviews would be concentrated in the town or city centre and self-completion questionnaires distributed elsewhere. A combination of techniques should be used at a small sample of stops so that a check for bias can be made.

Table 1 Advantages and Disadvantages of Different Public Transport Survey Methods

Method	Advantage	Potential Bias	Practical Difficulties	Response Rates	Appropriate Use
On-board face-to-face survey	Individual interviews of high quality	Fewer short-journey passengers, fewer peak passengers	Impossible to administer on crowded vehicles	From 10% on crowded vehicles to 90% on quiet services	Not busy buses, particularly in peak times.
At bus stop surveys; on-platform surveys	Individual interviews of high quality	Fewer 'runners', fewer peak passengers, no passengers who use those stops not surveyed	Unproductive, expensive if administered at all bus stops	From 5% at peak times to 60% inter-peak of those visiting survey bus stops	When there is no permission for on-board interviews, where it is acceptable that some stops are not surveyed. Where attitudes towards stop facilities required.
Entry/exit surveys at stations	Individual interviews of high quality	Fewer time-critical journeys, fewer 'runners', fewer peak passengers.	Very difficult to recruit at high passenger flows	From very few at peaks to 90% inter-peak	Not during morning peak hours at busy stations.
On-board self-completion survey	Highly productive	Self-selection biases. Some response bias against short-hop passengers	Poor quality and completeness of data, limited scope of data.	Up to 95% of survey forms returned but as low as 40% of questions answered	On highly crowded services, where only limited data required.
At bus stop or station self-completion survey	Productive	Self-selection biases, fewer time-critical journeys	Difficult to cover all bus stops on network, Poor quality and completeness of data, limited scope of data	Between 15% and 25% returned, with about 70% of questions answered	Where no permission for on-board survey, only limited data required, where it is acceptable that some stops are not surveyed.

5. Data for Specific Purposes

5.1 Introduction

5.1.1 This section provides advice on data requirements, and data sources that may be useful for particular modelling tasks or purposes, which is not a prescriptive list. There may be other emerging data sources not listed or discussed in this TAG unit that may be appropriate for specific purposes, when the general advice in Section 2 is followed.

5.2 Data Sources for Specific Purposes

5.2.1 Table 2 lists some useful data sources for specific but common modelling purposes. Many of the sources in this table are of use for a range of model inputs, but they also build on, and depend on other sources. For example, the National Trip End Model (NTEM) offers a valuable source of most of the data required to predict changes in trip ends, both trip productions based on household characteristics and trip attractions based on employment, as well as car availability forecasts.

Table 2 Data Sources for Specific Purposes

Modelling type	Data required	Data sources	Notes
Land use and trip end modelling	Land use	Census population and workplace population, ONS population data, VOA floorspace data, Local Authority planning data for employees, retail and commercial floorspace by zone, AddressBase, OS MasterMap Topography Layer	Employees and education places by zone enable calculation of zonal trip totals for doubly-constrained (to/from work) trip distribution. Some LUTI models need considerable detail for Social Economic Group (SEG), employment and floorspace data.
Land use and trip end modelling	Trip ends	NTEM	TEMPro is the tool to access NTEM outputs.
Demand modelling: trip generation	Population	NTEM, Census, ONS population data, Local Household Travel Survey	From wards or Census Output Areas amalgamated to zones: advisable to categorise by sex, employment status, income, household type and age (adult/child)
Demand modelling: trip generation	Employment	Census workplace population, VOA floorspace data, BRES,	

Modelling type	Data required	Data sources	Notes
		ASHE, ONS employment and labour market	
Demand modelling: trip generation	Households	NTEM, Census, Local Household Travel Survey	From wards or Census Output Areas amalgamated to zones
Demand modelling: trip generation	Education	DfE school census data	
Demand modelling: trip generation	Car ownership	NTEM / NATCOP, Census, Local Household Travel Survey, Household Expenditure Survey	Averaged across zones, or might be estimated from other socio-economic data
Demand modelling: trip generation	Socio-Economic Group	Census, Local Household Travel Survey, Household Expenditure Survey, Acorn, Mosaic	Not often used except in distinguishing workers from unemployed and retired, but some models categorise work travel by broad socio-economic groups (SEG), e.g. blue/white collar, and land-use models also require considerable SEG data.
Demand modelling	Car availability	Local Household Travel Survey, NTS	A proxy of the level of competition for a car between household members, can be estimated from household composition and license holding
Demand modelling	Licence holding	Local Household Travel Survey, Household Expenditure Survey, NTS	Provides increased resolution / refinement to car competition levels that helps distinguish mobility options better between household members
Demand modelling	Travel to zonal destinations by: Purpose, Mode, and Time of Day (Period)	RSI surveys, MND, ANPR surveys, Local Household Travel Survey, journey to work from Census, NTS	Full detail only available from specialised local surveys, by interview or questionnaire. NTS local samples are small, but values might be adjusted from wider NTS (e.g. split of purpose by time of day). Mode might include active modes and distinguish between car driver and passenger.
Demand modelling	Trip lengths	Local Household Travel Survey, NTS, journey to work Census, RSI surveys, Mobile Network Data	Not used directly since distance is specified by zonal structure but should be used in validation.

Modelling type	Data required	Data sources	Notes
Demand modelling	Vehicle operating costs	TAG unit A1.3 User and Provider Impacts, TAG Data Book	Perceived money costs of a vehicle journey are less than true average cost, and are assumed to be different for business and private travel.
Demand modelling: estimating costs	Vehicle occupancies	RSI surveys, Local Household Travel Survey, NTS, TAG Data Book	By purpose; often assumed from NTS or NTEM or TAG Data Book Table A1.3.3 - Sheet A1.3.3.
Demand modelling	Values of time	TAG Unit A1.3 User and Provider Impacts, TAG Data Book	Differ by purpose, and updated as appropriate. In principle, different values for behavioural modelling and standard values for appraisal
Base year matrix development (all modes)	Trip distribution data	MND, RSI/intercept surveys, RUDD, LENNON, MOIRA, ETM data, PT movement surveys for PT demand; NTS / household surveys	Refer to TAG Unit M2.2 Base Year Demand Matrix Development for advice, in particular Table 1 and Appendix B for matrix data sources, their strengths and limitations
Highway assignment modelling	Traffic counts	ATCs, MCCs, WebTRIS, TRICS, ANPR surveys	
Highway assignment modelling	Journey times	Moving Car Observer surveys, GPS tracking data, MNDANPR surveys, WebTRIS, Openrouteservice	Consider accuracy of data, choice of survey locations, need for information by vehicle type, cost of acquiring data, and need to increase the number of observations
Highway assignment modelling: designing zone system	Zone geography	ONS boundary data	
Highway assignment modelling	Network data	OS MasterMap Highways Network, OS Open Roads, OpenStreetMap	
Public Transport assignment modelling	Journey times	NRT, TNDS, BODS, NRE open data	
Public Transport assignment modelling: designing zone system	Zone geography	ONS boundary data, NaPTAN, OpenStreetMap	
Public Transport assignment modelling	Network data	NaPTAN, RUDD, NRT, TNDS, BODS	
Active mode modelling: network development	Network data	National Cycle Network, OpenStreetMap, Ordnance Survey	

Modelling type	Data required	Data sources	Notes
		MasterMap Highways Network	
Active mode modelling: demand development	Demand data	Propensity to Cycle Tool (PCT)[10]	
Activity-based demand modelling: activity sequences	Time spent undertaking activities by purpose, mode, and time of day	Local Household Travel Survey, NTS, Time Use Survey	A bespoke data collection exercise will generally be required

6. References

A303 project team and National Highways (2018), A303 Amesbury to Berwick Down, Combined Modelling and Appraisal Report. https://infrastructure.planninginspectorate.gov.uk/wp-content/ipc/uploads/projects/TR010025/TR010025-000451-7-5-ComMA.pdf. Accessed December 2022.

AECOM and PBA (2020), LATIS Lot 1 – Modelling Support, TN013 Derivation of Annualisation Factors (Car Driver + PT). https://www.transport.gov.scot/media/48744/website-tmfs18-technical-note-on-annualisation-factors.pdf. Accessed December 2022.

Atkins (2018), A2 Bean and Ebbsfleet Junction Improvements Economic Appraisal Package. https://assets.highwaysengland.co.uk/roads/road-projects/a2-bean-ebbsfleet-junction-improvements/Orders/I.7+Transport+Economics+Package.pdf. Accessed December 2022.

Fernandes, R., Krootjes, P. (2013). Navigating Away from Moving Observer Journey Times; Atkins' Application of TomTom Data to Evaluation. European Transport Conference, Frankfurt.

Gravett, N. and Mundaca, L. (2021). Assessing the economic benefits of active transport policy pathways: Opportunities from a local perspective. Transportation Research Interdisciplinary Perspective, Volume 11.

[10] This is a web-based tool for estimating cycling potential and corresponding health and CO2 benefits (down to the street level). The PCT covers travel behaviour data for commuting and travel to school, based on the 2011 Census and the 2011 school cycling Census.

O'Hare, P. and Jackson, P. (2020). Harnessing people movement data to shape future activity and people-centric designs. A case study of Salford Quays, European Transport Conference.

Steer and WSP (2019). Journeys per Season Ticket Study. Prepared for the Department for Transport, London, UK (https://www.gov.uk/government/publications/rail-journeys-per-ticket-study). Accessed December 2022.

Tolouei, R., Sayegh, A., Thornton, M.J. (2012), A Novel Approach to Developing LGV Trip Matrices for the Second Generation of Regional Traffic Models, European Transport Conference, Milan.

WSP (2020), Great Yarmouth Third River Crossing, TUBA Methodology Technical Note. https://www.norfolk.gov.uk/-/media/norfolk/downloads/roads-and-transport/3rd-river-crossing/supporting-documents/supporting-document-5-tuba-methodology-technical-note.pdf. Accessed December 2022.

7. Document Provenance

This unit is based on the previous TAG unit M1.2 (May 2020) and was substantially updated in spring 2024.